Extraction and annotation of 'location names'

UDC 81'322.2
DOI 10.18485/infotheca.2019.19.2.1

ABSTRACT: Introduced as part of the Message Understanding Conferences dedicated to information extraction, Named Entity extraction is a well-studied task in Natural Language Processing. The recognition and the categorisation of person names, location names, organisation names, etc., is regarded as a fundamental process for a wide variety of natural language processing applications dealing with content analysis and many research works are devoted to it, achieving very good results. One of our objectives is the identification and automatic (or semi-automatic) annotation of location names in order to apply the most appropriate information extraction methods. The main objective concerns the combination and interoperability between symbolic and statistical NLP (Natural Language Processing) methods (symbolic rules, machine learning, and data mining). Our work consisted of recognising named entities and in particular locations with Unitex, annotating them with Brat, and correcting them manually. The recall and accuracy rates are very encouraging but the question remains: What is a location name?

KEYWORDS: location names, locative complement, annotation, information extraction, Unitex.

PAPER SUBMITTED: 07 October 2019
PAPER ACCEPTED: 28 November 2019

Tita Kyriacopoulou
tita@u-pem.fr
Claude Martineau
claude.martineau@u-pem.fr

University of Paris-Est
Laboratoire d'Informatique
Gaspard-Monge, France

Markarit Vartampetian
markaritvar@gmail.com
Paris Nanterre University
Paris, France

1 Introduction

In this paper we will present a brief review of our research carried out on French location names. This work is done in the context of named entity extraction and annotation and, in particular, extracting and annotating locations in unstructured corpora. Over the last twenty years, considerable amount of work has emerged on this topic in the field of NLP (Natural Language Processing), MUC evaluations[1] (1987-1998), ACE program (Doddington et al., 2004), NIST’s[2], and ATALA’s[3] work. The main difficulty lies, no longer, in the named entity recognition, but in the named entity disambiguation and their correct interpretation; for instance, distinguishing the organisation *aéroport Charles de Gaulle* (Charles de Gaulle Airport) in *L’aéroport Charles de Gaulle va ouvrir ses portes à un nouveau terminal* (Charles de Gaulle Airport will open its doors to a new terminal) from the location *On arrive à l’aéroport Charles de Gaulle dans 15 minutes* (We arrive at Charles de Gaulle airport in 15 minutes). That is, the analysis must assign a category according to the context. Despite the work done so far, research on named entities still plays a central role in NLP and especially in the process of disambiguation of corpora.

It is worth nothing that the choice of Location Names was made because of the University Gustave Eiffel – the University of Marne-la-Vallée is part of it – which will be founded by January 1st 2020 and whose topic will be ‘the City’.

First, we will present our research data, based on corpora works, and then the different categories of location names. Next, we will discuss the problem of ambiguity and we will try to make a clear distinction between location names and locative complements. Besides, we will describe our annotation method of location names and we will compare the results with those of the semi-automatic annotation[4] by use of the open source tool Gemini.

2 Collecting data

Within the framework of our research, the language resources consist of a combination of lexicons available in Unitex/GramLab,[5] and of freely avail-

[1] Message Understanding Conference

[2] National Institute of Standards and Technology

[3] Association pour le Traitement Automatique des Langues (French Association for Natural Language Processing)

[4] ‘Nous entendons une annotation ... ’(Brando et al., 2016)

[5] Unitex/GramLab

able corpora. These corpora were preprocessed and analysed in Unitex[6] and annotated in XML format. In particular, we have worked on four corpora[7]:

- Corpus de la presse
  Press corpus
- Le Tour du monde en 80 jours
  Around the world in eighty days
- 20.000 lieues sous les mers
  Twenty thousand leagues under the sea
- Extraits de 2001 à 2012 du Monde Diplomatique
  Monde Diplomatique extracts from 2000 to 2012

Table 1 lists the main features of these corpora. The parentheses in tokens and simple forms indicate the number of different forms.

Corpus	Press . . .	Around . . .	Twenty . . .	Monde . . .
sentence delimiters	1,756	3,652	7,349	8,387
tokens	100,642 (8,467)	165,239 (9,452)	339,425 (15,301)	479,808 (22,952)
simple forms	43,664 (8,404)	71,859 (9,422)	149,007 (15,268)	205,177 (22,860)
digits	3,742 (10)	438 (10)	1,047 (10)	11,892 (10)
simple-word entries	10,082	13,229	20,654	199,417
compound entries	2,039	2,099	3,123	5,586
unknown simple words	1,405	3,156	2,161	4,877

Table 1. Overview of corpora

It should be noted that the corpora do not consist of only narrative content. Some sentences constitute simple utterances of terms, recalling the inventory process:

[6] For the sake of conciseness we will be using Unitex instead of Unitex/GramLab in the rest of the article.

[7] The second and third corpora are novels written by Jules Verne.

(1) îles Salcette, Colaba, Éléphanta, Butcher
Salcette, Colaba, Éléphanta, Butcher islands

(2) au large des Pomotou, des Marquises, des Sandwich
off the coastline of Pomotou, Marquises, Sandwich

In example (1) note that, because of the enumerative process, the trigger word *îles* (*islands*) is present only in front of the first name, whereas in example (2) the same trigger word is absent. The enumerative sequence is introduced by *au large*. Besides, we notice that the words *Marquises* and *Sandwich* are ambiguous.

Similarly, forms with connectors (and, or, etc.) can be complex to analyze.

(3) cratères de l'Érébus et du Terror
Erebus and Terror craters

(4) affluents ou sous-affluents de la Little-Blue-river
tributaries or sub-tributaries of the Little-Blue-river

(5) côtes arabiques du Mahrah et de l'Hadramant
arabian coasts of Mahrah and Hadramant

In example (3), in addition to a coordination of names of volcanoes, there is a trigger word which is not the word *volcan* (volcano), but the word *cratère* (crater), which complicates even further the recognition. The example (4) contains a disjunction *affluents ou sous-affluents* (tributaries or sub-tributaries) which consists of trigger words and indicates that the following is a hydronym (river, stream) as in the phrase *les affluents de la Seine* (tributaries of the Seine). In *Little-Blue-river* the status of the hydronym is indicated by the word *river*. But in this case, *affluents* and *sous affluents* are not named entities, but entities of type hydronym in relation to the named entity *Little-Blue-river*.

Finally, example (5) constitutes yet another coordination for location names. However, the precise location is defined both by the noun *côtes* (coasts) and the toponymic adjective *arabiques* (arabian).

As the extraction systems appear to have been trained more in continuous texts rather than in non-continuous texts, it is difficult for them to deal with this non-narrative content.

The absence of context may occasionally lead to false recognition. For instance, in example (6) *aéroport de Montréal* clearly indicates a location. However, in example (7) we notice that the context *porte-parole* (spokeswoman) has a double impact on the interpretation. On the one hand, it confers on the *aéroport de Montréal* an organisation 'role' and, on the other hand, it refers to *porte-parole* (the spokeswoman), but not to the organisation itself.

(6) Quand on arrive à l'aéroport de Montréal, ...
When we arrive at Montreal airport ...

(7) La porte-parole de l'aéroport de Montréal assure, pour sa part, ...
The spokeswoman for Montreal airport ensures, ...

3 Location Names

Regarding the named entities and, in particular, locations names, notable works have emerged on typologies, which allow us to define what we aim to recognise and extract. Several typologies have been proposed and they differ in the categories as well as in the structure of the elements and even in the typology itself. Sometimes, they also differ fundamentally in the definition of the notion of named entities

For instance in TEI[8] the sequence *à 20 km au nord de Paris* is annotated as follows:

```
<placeName>
   <measure>20 km</measure>
   <offset>au nord de</offset>
   <settlement type="city">Paris</settlement>
</placeName>,
```

whereas in ISTEX[9] it is annotated as:

```
<placeName>
    Paris
</placeName>.
```

3.1 Categories

According to the typologies in use, the number of named entity categories varies. However, certain categories are present in most typologies:

[8] TEI

[9] ISTEX

- Names of persons or saints: 'Alexander the Great', 'Saint Ambroise', 'Constantine the Great', 'Saint Demetrius.'
- Location names[10] (names of countries, cities, regions, etc.): 'Eastern Europe,'. Note that institutions, organisations and even events act as location names in some contexts, for instance *Aller au Salon du Livre de la Havane* (cf. Section 5).
- Names of mountains (oronyms) or hydronyms: 'Saint Basil Lake','the Black Sea.'
- Organizations: 'Agricultural Bank', 'Orsay Museum', 'Ministère de l'Agriculture'
- Authorial texts: 'Green Paper,' 'the National Anthem'
- Events: 'The French Revolution' 'The Olympic Games','Le 11 Septembre'

According to (Chinchor, 1998), named entities include proper nouns but also time expressions, such as 'Holy Week', 'Nautical Week', 'Black Friday'.

The task of named entity extraction consists of automatically recognising the NE in corpus, extracting and classifying them into categories such as *Person*, *Location*, *Organization.* As indicated by (Denis and Sagot, 2012), we can distinguish two ways of identifying an entity, either intrinsic where *France* denotes a place, or contextual, as in *France signed the treaty*, where *France* can be recognised as an organisation.

As it is very aptly described in (Hengchen et al., 2015), during the analysis of a corpus, various questions arise regarding the classification of named entities of type location. Even though the categories are well defined (at first, a location is not an organisation and vice versa), their rigidity requires the researcher to make subjective choices. *Paris* is used in a geographical sense (town), but Paris may also indicate a name, thus it can be logically categorised as location *LOC* and *PERSON.* But what about the term *Paris* in contexts referring to the town of Paris as an organisation *ORG*, e.g., in *Paris va organiser les jeux olympiques* (Paris will organise the Olympic Games)? Therefore, a term which has a fixed geographical location by nature, but represents a commercial enterprise, is it to be considered as a location or an organisation? In fact, this is the key question in this article

[10] In the Prolex dictionary provided with the Unitex distribution, all the location names are marked as toponym with a more specific feature (City, Country, Hydronym, etc.)

and to which we intend to answer.

Furthermore, a great number of named entities, especially organisations and/or locations, may appear as initialisms or acronyms: (Université Paris-Est Marne-la-Vallée UPEM, Made in USA), which complicates the interpretation. *CE*, for example, can indicate *conseil d'Etat* (Council of State), *conseil de l'Europe* (Council of Europe), *Comité d'Entreprise* (works council), etc. In Wikipedia there are more than ten interpretations for *CE*.

Named entity variations were also present in our corpora, e.g., *Muséum de Paris* which can be designated by different variations :

(8) Muséum National d'Histoire Naturelle de Paris
National Natural History Museum of Paris

(9) Muséum National d'Histoire Naturelle.
National Natural History Museum

(10) Muséum National de Paris
National Museum of Paris

(11) Muséum d'Histoire Naturelle
Natural History Museum

(12) Muséum de Paris
Museum of Paris

(13) MNHN
NNHM

Deleting certain terms influences considerably the accuracy of recognition. Moreover, the English translation generates additional ambiguity. In French, the word *Muséum* designates a scientific museum devoted to natural sciences. For other domains the word *Musée* is used more (e.g. Musée du Louvre (Louvre Museum), Musée Guimet (Guimet Museum), Musée des Arts Décoratifs (Museum of Decorative Arts), etc.). Hence, in French, *d'Histoire Naturelle* can be deleted without resulting in ambiguities, since the domain is indicated by the word *Muséum*. The location Paris is indicated either by the presence of the word *Paris* or by the adjective *National* because the other "Muséums" that exist (in Le Havre, Grenoble, La Rochelle, etc.) are not qualified as nationals. This example illustrates the fact that the accuracy of the recognition depends on the terms that are erased and that this erasability is language dependent. In English, *Natural History* cannot be erased without losing information.

4 What is a location name

Location names (according to typology *LOC*, *placeName*, *Loc.admi*, etc.) as other named entities may be more or less ambiguous and may depend on corpus but in a given context they can act as metonymy. Sometimes even we have translation problems, especially when a location name refers to different locations. For instance, *London* denotes a town in Great Britain as well as in Canada. However, the French translation differs: *Londres* designates the capital of Great Britain, but *London* refers to the town across the Atlantic.

Considering the following sentences (14) et (15):

(14) Marie va à Paris
Mary is going to Paris

(15) Paris va organiser les jeux olympiques en 2024
Paris will organise the Olympic Games in 2024

the town of *Paris* denotes a location in (14), but an administrative authority, namely an organisation, in (15). Hence, defining a location remains an acute problem.

(16) Marie habite Rue de Paris
Marie lives in Rue de Paris

Likewise, in (16) the sequence *Rue de Paris* does not indicate a precise location since *Rues de Paris* exist in Lille, Nice, and other cities including Paris. However, in (17) *Paris* denotes a sports team, and probably a football team.

(17) Paris a battu Lille 2-0
Paris beat Lille 2-0

Finally, it is known that location names have always inspired first name choices. Thus, *France* does not only constitute a country, but also a first name.

These issues are partly solved by Ester 2 initiative,[11] which proposes named entity sub-categories in addition to the traditional categories, as it is shown in the following examples:

[11] Ester 2 initiative

(18) Je suis stationné à côté de la <ent type ="loc.fac">mairie de Paris</ent>.
I am parked next to the city hall of Paris

(19) La course à la <ent type ="org.pol">mairie de Paris</ent> a commencé.
Paris mayor's race has begun

(20) Je me suis fait opérer à l'<ent type ="org.non-profit">hôpital Necker</ent>.
I had a surgery at Necker Hospital

Thus, *la ville de Paris* would be *LOC.admi*, namely a *LOC* entity type related to an administrative authority. Nevertheless, this type of classification is rarely exhaustive and even less integrated in information extraction systems which extract named entities without relating them to a given context.

Afterwards, we shall present a number of questions that need to be raised.

4.1 Named Entity or Extended Named Entity

Do we intend to recognise a Named Entity or an Extended Named Entity? Gaio and Moncla (2017) have used the concept of Extended Named Entity (ENE). Based on Jonasson's definition, an ENE refers to an entity built with a proper name, e.g., *Rue de Paris*, and possibly composed of one or more concepts, e.g., *La maire de Paris.*

We believe that when we have to extract and annotate locations, it is evident that we should recognise the ENE rather than the NE since in (21) *La maire de Paris* denotes a person.

(21) La maire de Paris
The mayoress of Paris

However, regarding the ENE, the referential nature of named entities can generate difficulties and pose problems that are not solved yet. For instance, Sagot (Sagot et al., 2012) argues that this situation is found in annotation cases of university names. Therefore, according to Sagot, *université*[12] *de Marne-la-Vallée* denotes a university located in Marne-la-Vallée and we should only annotate the town Marne-la-Vallée, whereas *Université*

[12] The word is written without a capital letter.

de Marne-la-Vallée refers directly to the organisation that the university constitutes and, as a result, we should annotate the whole entity as an organisation. Similarly, regarding the example of *Université de Montpellier*, since a unique organisation corresponding to this term does not exist, only *Montpellier* should be annotated as a town.

Hence, the above examples demonstrate that in some cases we need to recognise an ENE, e.g., *maire de Paris*, *rue de Paris*, *Université de Nantes*, whereas in other cases it is sufficient to recognise only an NE, e.g., *ville de Paris*. From our perspective, it is almost impossible to distinguish the trigger words that are part of the entity from those who do not without leaving out the trigger words that are part of the entity, e.g., in *Mer Morte* (Dead Sea). The solution proposed by Unitex with the use of graphs and advanced functions (use of contexts, weights, variables, etc.) seems – for us – to be the best compromise, which allows us to choose the annotation according to the needs of the applications and of the corpora.

From our perspective, this type of precision is necessary as we are interested in the context and in the exact and precise named entity annotation, with a view to using it in machine learning systems.

In this research, we decided to recognise the ENE even if *université de Paris* is not a unique reference and/or does not refer to a different era.

5 Ambiguities

As we have shown various examples of ambiguities concerning location names, we could say in brief, that the most problematic cases are:

1. distinguishing a location from a locative complement,
2. distinguishing a location or a locative complement from an organisation, and
3. distinguishing *Loc.admi* from *Loc.Person.*

The first two cases are related, among others, to the syntactic description, thus considering the named entity not as an isolated sequence (simple or complex), but as an element of a basic sentence. In bibliography, we often refer to the context, but the notion of context is more vague and almost never explicit.

For the first case, it is important to specify that we consider as a location every toponym generally used in a geographical sense – see example (22) –

whereas a locative complement answers to the question *where* – see examples (23) and (24).

(22) Paris est une belle ville
Paris is a beautiful city

(23) Claire va à Paris
Claire goes to Paris

(24) Claire dort à Paris
Claire sleeps in Paris

To push our analysis further consider the following examples:

(25) Claire se repose à Paris / Claire va à Paris
Claire rests in Paris / Claire goes to Paris

(26) Claire se repose. / *Claire va.
Claire rests. / *Claire goes.

(27) À Paris, Claire se repose. / *À Paris, Claire va.
In Paris, Claire rests / To Paris, Claire goes

In these examples, we observe that there are specificities between the verbal predicate *se reposer* (to rest) and *aller* (to go) and yet in both cases the locative complement *Paris* answers the question *where*. These differences emerge from the fact that in the case of the verb *se reposer à Paris* is a modifier whereas in the case of the verb *aller* is an argument. It is noteworthy, that in English this distinction results in two different translations (*to Paris*, *in Paris*). In order to make this distinction a syntactico-semantic analyser is necessary.

Regarding the second case, we consider as organisation names all references to an organisation; political, educational, financial, religious, associative, etc., which are annotated as *Loc.admi* in Ester 2. In this case, syntactic analysis is able to give solutions:

(28) La ville de Paris va organiser les jeux olympiques en 2024/Paris va organiser les jeux olympiques en 2024

The city of Paris will organise the Olympic Games in 2024/Paris will organise the Olympic Games in 2024

(29) Les jeux olympiques auront lieux à la ville de Paris / Les jeux olympiques auront lieux à Paris
Olympic Games will take place in the city of Paris/Olympic Games will take place in Paris

Thus, by applying the same tests as in examples (28) and (29) we observe that in example (29) the segments *ville de Paris* / *Paris* answer to the question *where* and, as a result, they constitute locative complements, which are considered as arguments by linguists (Gross, 1996) because a sentence like (30) is not acceptable.

(30) *Les jeux olympiques auront lieu
*Olympic Games will take place

However, not all ambiguities that we have encountered in our corpus can be solved by syntactic analysis.

Regarding the third case, disambiguating a *Loc.admi* (a *Loc* entity type related to an administrative authority) from a *Loc.person* (a *Loc* entity type related to a person) is more complicated. Thus, between :

(31) Paris organise les jeux olympiques de 2024
Paris will organise Olympic Games in 2024

(32) Paris organise une grande fête pour le 14 juillet
Paris will organise a great feast for the 14th of July

the syntactico-semantic analysis is not able to distinguish the town of *Paris* from a person. In example (31), extralinguistic information is necessary for a machine to know that a person cannot organise an event of this type. Hence, this sentence must be modified as *La ville de Paris* organise les jeux olympiques de 2024. Still, example (32) remains ambiguous even for a human being.

Note that all the ambiguities cannot be solved with syntactico-semantic analysis.

6 Methodology

In order to deal with the above problems, we used the following method: First, we created graphs using Unitex and we proceeded to an automatic annotation in XML format. Then, we imported the annotated corpora using a script to Brat (Brat rapid annotation tool),[13] we validated the annotated corpora, and, finally, we corrected the annotations manually. Brat is a tool that allows us to visualise in a browser (Mozilla, Opera, etc.) the annotated corpora by highlighting the recognised patterns with a color associated with its named entity type.

The grammar of Figure 1 allows us to annotate the text by adding tags *Lieu* (<Lieu> </Lieu>) to the recognised patterns. At the bottom of the graph, we notice two paths that allow us to recognise ambiguous patterns containing a location name without tagging it. Figure 2 shows an extract of the results in concordance form. Figure 3 shows certain concordances in which the tag *Lieu* contains a precise type (country, region, hydronym, etc.). Figure 4 shows some sequences recognized by the grammar for purposes of disambiguation and therefore not annotated.

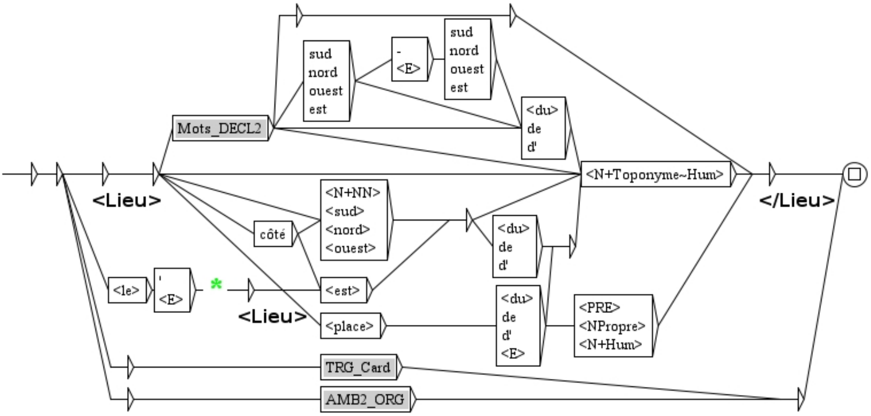


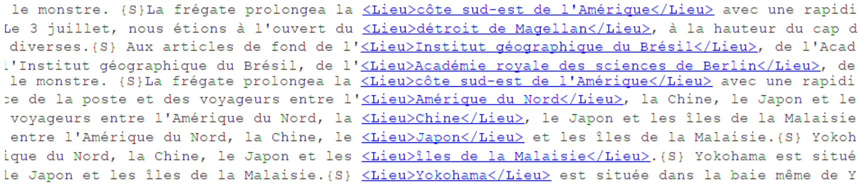
Figure 1. Grammar of location names annotation

.

We proceeded with two types of manual annotation. The first one aims at annotating arguments, whereas the second takes into account all location names. The grammars that have been constructed for the automatic annotation by now perform correctly for the second type of annotation and will

[13] Brat

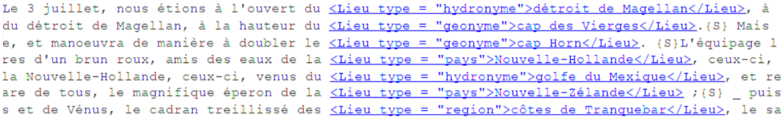
be improved so as to be able to recognise only the arguments. As a result, during our evaluation of the results, we compared the manual annotation with all locations.



```
 le monstre. {S}La frégate prolongea la <Lieu>côte sud-est de l'Amérique</Lieu> avec une rapidi
Le 3 juillet, nous étions à l'ouvert du <Lieu>détroit de Magellan</Lieu>, à la hauteur du cap d
 diverses.{S} Aux articles de fond de l'<Lieu>Institut géographique du Brésil</Lieu>, de l'Acad
l'Institut géographique du Brésil, de l'<Lieu>Académie royale des sciences de Berlin</Lieu>, de
 le monstre. {S}La frégate prolongea la <Lieu>côte sud-est de l'Amérique</Lieu> avec une rapidi
ce de la poste et des voyageurs entre l'<Lieu>Amérique du Nord</Lieu>, la Chine, le Japon et le
 voyageurs entre l'Amérique du Nord, la <Lieu>Chine</Lieu>, le Japon et les îles de la Malaisie
 entre l'Amérique du Nord, la Chine, le <Lieu>Japon</Lieu> et les îles de la Malaisie.{S} Yokoh
ique du Nord, la Chine, le Japon et les <Lieu>îles de la Malaisie</Lieu>.{S} Yokohama est situé
le Japon et les îles de la Malaisie.{S} <Lieu>Yokohama</Lieu> est située dans la baie même de Y
```

Figure 2. Concordances of location names

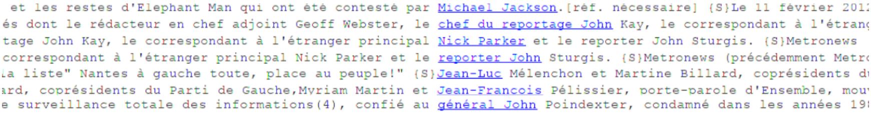
.



```
Le 3 juillet, nous étions à l'ouvert du <Lieu type = "hydronyme">détroit de Magellan</Lieu>, à
du détroit de Magellan, à la hauteur du <Lieu type = "geonyme">cap des Vierges</Lieu>.{S} Mais
e, et manoeuvra de manière à doubler le <Lieu type = "geonyme">cap Horn</Lieu>. {S}L'équipage l
res d'un brun roux, amis des eaux de la <Lieu type = "pays">Nouvelle-Hollande</Lieu>, ceux-ci,
la Nouvelle-Hollande, ceux-ci, venus du <Lieu type = "hydronyme">golfe du Mexique</Lieu>, et re
are de tous, le magnifique éperon de la <Lieu type = "pays">Nouvelle-Zélande</Lieu> ;{S} _ puis
s et de Vénus, le cadran treillissé des <Lieu type = "region">côtes de Tranquebar</Lieu>, le sa
```

Figure 3. Concordances of location names with types

.



```
 et les restes d'Elephant Man qui ont été contesté par Michael Jackson.[réf. nécessaire] {S}Le 11 février 201
és dont le rédacteur en chef adjoint Geoff Webster, le chef du reportage John Kay, le correspondant à l'étran
tage John Kay, le correspondant à l'étranger principal Nick Parker et le reporter John Sturgis. {S}Metronews
correspondant à l'étranger principal Nick Parker et le reporter John Sturgis. {S}Metronews (précédemment Metr
la liste" Nantes à gauche toute, place au peuple!" {S}Jean-Luc Mélenchon et Martine Billard, coprésidents d
ard, coprésidents du Parti de Gauche,Myriam Martin et Jean-Francois Pélissier, porte-parole d'Ensemble, mou
e surveillance totale des informations(4), confié au général John Poindexter, condamné dans les années 19
```

Figure 4. Concordances of recognized sequences not tagged as location names

.

7 Comparing automatic annotations with semi-automatic annotations

In this section, we compare the automatic annotation produced by Unitex with the semi-automatic annotation (manual correction of the automatic annotation by an experienced linguist) This task was done automatically by use of the Gemini[14] tool, developed in LIGM,[15] which allows us to calculate the *Precision*, the *Recall*, and the *F-measure* values:

[14] Gemini source on Github

[15] Laboratoire d’Informatique Gaspard-Monge

$$Precision = \frac{number\ of\ correctly\ matched\ location\ names}{number\ of\ terms\ matched\ as\ location\ names}$$

$$Recall = \frac{number\ of\ correctly\ matched\ location\ names}{number\ of\ actual\ location\ names}$$

$$F = 2 * \frac{precision * recall}{precision + recall}$$

The results obtained from our corpora are listed in Table 2. We notice that in the newspaper corpus the recall has increased. This can occur because of the various ambiguities in that corpus. As the corpus Monde Diplomatique is more homogeneous, it is affected less by this phenomenon.

Corpus	Press corpus	Around the ...	Twenty Thou ...	Monde Diplo ...
Precision	0.24403422	0.74358974	0.78455056	0.51353696
Recall	0.89144737	0.72432932	0.68691588	0.47702724
F-measure	0.38317427	0.73383317	0.73249409	0.49460927

Table 2. Results generated by the Gemini comparison tool

The following constitute true (positive, negative) and false (positive, negative) examples issued from our corpora and the application of grammars.

- True positive :
  - ... on apprit qu'un steamer de la ligne de San Francisco de Californie à Shangaï avait revu l'animal, trois semaines auparavant, dans les ***mers septentrionales du Pacifique*** ...
    ... word came that a steamer on the San Francisco line sailing from California to Shanghai, had sighted the animal again, three weeks before in the ***northerly seas of the Pacific*** ...
  - Mais, en attendant, il me fallait chercher ce narwal dans le ***nord de l'océan Pacifique***
    But in the meantime, I would have to search this narwhale in the ***northern Pacific Ocean***

- Tout deux se rendront au ***lycée Paul Cézanne d'Aix-en-Provence*** qui propose une section expérimentale de langues et cultures méditerranéennes.
  Both will go to ***Paul Cézanne d'Aix-en-Provence high school*** which proposes an experimental section of Mediterranean languages and cultures

- ... plusieurs instituts dédiés à l'enseignement du droit international virent le jour en Amérique et en Europe, tel l'***Institut universitaire de hautes études internationales de Genève*** ...
  ... several institutes dedicated to international legal studies have emerged in America and in Europe, such as the ***Graduate Institute of International Studies in Geneva*** ...

- Une centaine de manifestants ukrainiens ont pénétré samedi dans le ***principal bâtiment du ministère de l'Énergie*** ...
  About one hundred Ukrainian protesters entered the ***main building of the Ministry of Energy*** on Saturday ...

- C'est à 20h45mn qu'il a foulé le ***tarmac de l'aéroport international Léopold Sédar Senghor de Dakar***, à bord de la compagnie d'Air France.
  It was 20:45 when he crossed the ***tarmac of Léopold Sédar Senghor International Airport in Dakar***, onboard of the Air France company.

– True negative :

- ... ancien chef de l'ANC, devenu président de l'***Afrique du Sud***
  ... former leader of the ANC, became president of ***South Africa*** ...

- Cette ambiguïté, ou plutôt cette confusion, autour du mot « science » permet au sociologue Alain ***Touraine*** de disculper ...
  This ambiguity, or rather this confusion, around the word "science" allows the sociologist Alain ***Touraine*** to exonerate ...

- C'étaient des portraits, des portraits de ces grands hommes historiques dont l'existence n'a été qu'un perpétuel dévouement à une grande idée humaine, Kosciusko, le héros tombé au cri de Finis Polonioe, Botzaris, le Léonidas de la Grèce moderne, O'Connell, le défenseur de l'Irlande, ***Washington***, le fondateur de l'Union américaine ...
  They were portraits of these great men of history who had spent

their lives in perpetual devotion to a great human ideal, Kosciusko, the hero whose dying words had been Finis Poloniae, Botzaris, the Leonidas of modern Greece, O'Connell, Ireland's defender, ***Washington***, founder of the American Union ...

- Elles furent ensuite nommées Malouines, au commencement du dix-huitième siècle par des pêcheurs de ***Saint-Malo*** ...
  at the beginning of the 18th century, they were named the Malouines by fishermen from ***Saint-Malo*** ...
- Si aujourd'hui il est difficile d'imaginer la firme de Cupertino passer à ***4 Go de RAM*** un de ses mobiles ...
  If today it is difficult to imagine the Cupertino firm move one of his mobiles to ***4GB RAM*** ...

– False positive :

- ***Allons*** donc !
  Let's go!
- « ***Viens*** là »
  Come on!
- Et si l'on en croit la leçon du beau film Little ***Sénégal***
  And if we believe the lesson of the beautiful film Little ***Senegal***
- ***Paris 2002***
- Vêtus comme des hérauts du ***Moyen*** Age
  Dressed as heralds of the ***Middle*** Ages
- ***Né*** dans l'ombre du pouvoir
  Born in the shadow of power
- ***Mars***

– False negative :

- New Zealand's Role in the International Spy Network, Craig Potton Publishing, ***Nelson***, Nouvelle-Zélande, 1996.
- ... de ses voyages et des trois petites Nahîla nées à ***Deir-el-Assad*** ...
  ... about his travels and the three little Nahilas born at ***Deir-el-Assad*** ...

- ... et exploitant à ***Marly-le-Roi*** (78), Trappes (78), Les Clayes-sous-Bois (78), Asnières (92), Nanterre (92), Boussy-Saint-Antoine (91) et Epernay (51)
  ... and operator in ***Marly-le-Roi*** (78), Trappes (78), Les Clayes-sous-Bois (78), Asnières (92), Nanterre (92), Boussy-Saint-Antoine (91) et Epernay (51)
- Il ne s'est pas opposé à l'installation des prisonniers d'Al-Qaida sur la ***base militaire de Guantanamo Bay***
  He did not oppose to the installation of Al-Qaida prisoners at the ***Guantanamo Bay military base***
- L'auteur est née en 1968 à ***Niodor***...
  The author was born in 1968 in ***Niodor*** ...

8 Conclusion

Our first goal was the annotation of location names based on four corpora (469,707 words in total) by use of Unitex – and of Brat regarding the manual correction – in order to, ultimately, test the GEMINI tool, which is able to compare an automatic annotation with a semi-automatic one (automatic annotation manually corrected by an experienced linguist). But we have readily been confronted with the complexity of this named entity type, to the point of asking the question *what is a location name*. Our main contribution in this article was to improve the recognition of these named entity types without resolving all the ambiguities presented in our corpora. While most tools do not recognise a location name as a type *LOC* named entity if it refers to another type of entity, such as organisation, person or event, we have tried to disambiguate this type of information and annotate it according to the context and not to the ambiguities. We strongly believe that by integrating a syntactic analyser in our system, the results will ameliorate. However, we need to annotate more corpora and process more languages in order to achieve satisfactory results.

References

Brando, Carmen, Nathalie Abadie and Francesca Frontini. "Évaluation de la qualité des sources du Web de Données pour la résolution d'entités nommées". *Ingénierie des Systèmes d'Information* (2016), numéro spécial 'Web de données : publication, liage et capitalisation'

Chinchor, Nancy A. “Proceedings of the Seventh Message Understanding Conference (MUC-7) Named Entity Task Definition”. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, Fairfax, VA, 1998. http://acl.ldc.upenn.edu/muc7/ne_task.html, version 3.5, http://www.itl.nist.gov/iaui/894.02/related_projects/muc/

Denis, Pascal and Benoît Sagot. “Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging”. *Language Resources and Evaluation* Vol. 46, no. 4 (2012): 721–736. https://hal.inria.fr/inria-00614819

Doddington, George, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel et al. “The Automatic Content Extraction (ACE) program-tasks, data, and evaluation”. *Proceedings of LREC* Vol. 2 (2004)

Gaio, Mauro and Ludovic Moncla. “Extended Named Entity Recognition Using Finite-State Transducers: An Application To Place Names”. In *The Ninth International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2017)*. Nice, France, 2017. https://hal.archives-ouvertes.fr/hal-01492994

Gross, Maurice. *GRAMMAIRE transformationnelle DU FRANÇAIS: 1 - Syntaxe du verbe*, Cantilène, 1996

Hengchen, Simon, Seth van Hooland, Ruben Verborgh and Max De Wilde. “L’extraction d’entités nommées : une opportunité pour le secteur culturel ?”. *Information, Données et Documents* Vol. 52, no. 2 (2015): 70–79

Sagot, Benoît, Marion Richard and Rosa Stern. “Annotation référentielle du Corpus Arboré de Paris 7 en entités nommées”. In *Traitement Automatique des Langues Naturelles (TALN)*, Antoniadis, Georges, Hervé Blanchon and Gilles Sérasset, *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, Vol. 2 - TALN, Grenoble, France, 2012. https://hal.inria.fr/hal-00703108